

Interconnessioni tra classificazioni scientifiche e classificazioni generali nel mondo digitale

Antonella De Robbio

e-mail: derobbio@math.unipd.it

Biblioteca del Seminario Matematico, Università di Padova

Dario Maguolo

e-mail: dario@math.unipd.it

Abstract

Il lavoro che viene qui presentato si riferisce ai problemi di comunicazione fra i mondi specialistici della scienza e della tecnologia da una parte e il mondo delle biblioteche dall'altra in relazione alle classificazioni disciplinari specifiche da una parte e le classificazioni generali utilizzate nelle biblioteche, in particolare la Classificazione Decimale Dewey (CDD) dall'altra.

Dopo aver richiamato alcune nozioni generali sulla struttura, la dinamica e le mutue interferenze delle classificazioni semantiche, si espongono le linee essenziali di una sinossi diacronica delle classificazioni specifiche per le aree della matematica, dell'informatica, dell'ingegneria, della fisica e dell'astronomia.

Di seguito si passa in rassegna lo sviluppo della CDD in queste aree, fino alla revisione recentemente proposta della sezione 510 della CDD (matematica, con l'esclusione dell'informatica teorica e della fisica matematica) nella prospettiva della prossima edizione (22.) della CDD.

Questo articolo è, in parte, una rielaborazione in lingua italiana del lavoro dal titolo "Mathematics Subject Classification and related Classifications in the Digital World", che appare nei Proceedings della conferenza internazionale "Crimea2001" tenutasi a Sudak, in Ukraina, dal 9 al 17 giugno 2001¹. In quella occasione è stato presentato il nuovo sito La pagina delle classificazioni scientifiche², che raccoglie i prodotti del lavoro svolto finora.

Introduzione

¹ Eighth International Conference "Crimea 2001" "Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation". Section 4: Electronic Libraries. Sudak, Ukraine, 9-17 giugno 2001. Lavoro presentato da Antonella De Robbio, Dario Maguolo e Alberto Marini
<<http://www.gpntb.ru/win/inter-events/crimea2001/tom/sec4/Doc23.HTML>>

² <<http://www.math.unipd.it/~biblio/math/index.html>>

Il lavoro che viene qui presentato si riferisce ai problemi di comunicazione fra i mondi specialistici della scienza e della tecnologia da una parte e il mondo delle biblioteche dall'altra, evidenziati dagli scarti a livello linguistico, strutturale e semantico fra le classificazioni disciplinari specifiche e le classificazioni generali utilizzate nelle biblioteche, in particolare la Classificazione Decimale Dewey (CDD).

La CDD è largamente usata in tutto il mondo non solo per la collocazione dei libri nelle biblioteche, specialmente nelle biblioteche pubbliche, scolastiche e generali di università, ma anche per l'indicizzazione semantica in strumenti generali per la ricerca e l'accesso in linea a documenti, quali database bibliografici, cataloghi online di biblioteche (OPAC), biblioteche digitali, motori di ricerca nel Web.

In questo contesto risulta assai utile mettere a punto strumenti, ausiliari agli OPAC, ma non solo, che mettano in colloquio differenti ambiti disciplinari e classificazioni generali, come per esempio la Dewey, entro un'infrastruttura che integri i vari contenitori informativi (OPAC, banche dati, archivi di documenti, ecc.)

La creazione di strumenti per la connessione tra schemi diversi prevede una sequenza di fasi operative che richiedono:

1. la messa a punto di una metodologia per interconnettere le classificazioni tramite corrispondenze lessicali nelle descrizioni testuali dei codici di classificazione
2. l'applicazione di strumenti software per lo sviluppo di presentazioni ipertestuali di classificazioni da file sorgenti sequenziali.
3. la costruzione di complessi di pagine Web che realizzano visualizzazioni ipertestuali, singole o combinate, di classificazioni

Sono stati sviluppati dal nostro gruppo di lavoro strumenti software per realizzare diverse modalità ipertestuali di visualizzazione di classificazioni; tali strumenti sono stati messi a punto con riferimento, oltre che alla CDD, a classificazioni specializzate nei campi della matematica, della fisica e delle scienze e tecnologie correlate. Gli schemi interessati sono Mathematics Subject Classification (MSC) e Zentralblatt für Didaktik der Mathematik (ZdM) Classification Scheme per la matematica, Physics and Astronomy Classification Scheme (PACS) per la fisica, Association for Computing Machinery - Computing Classification System (ACM-CCS) per l'informatica.

Questo articolo è, in parte, una rielaborazione in lingua italiana dell'articolo dal titolo "Mathematics Subject Classification and related Classifications in the Digital World", che appare nei Proceedings della conferenza internazionale "Crimea2001" tenutasi a Sudak, in Ukraina, dal 9 al 17 giugno 2001³. In quella occasione è stato presentato il nuovo sito *La pagina delle classificazioni scientifiche*⁴, che raccoglie i prodotti del lavoro svolto finora.

Il sito, disponibile anche in versione inglese, è strutturato in tre sezioni:

³ Eighth International Conference "Crimea 2001" "Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation". Section 4: Electronic Libraries. Sudak, Ukraine, 9-17 giugno 2001. Lavoro presentato da Antonella De Robbio, Dario Maguolo e Alberto Marini
<<http://www.gpntb.ru/win/inter-events/crimea2001/tom/sec4/Doc23.HTML>>

⁴ <<http://www.math.unipd.it/~biblio/math/index.html>>

1. *La pagina della classificazione matematica*, contenente i sei strumenti descritti in un precedente articolo pubblicato su Biblioteche Oggi nel maggio del 2000⁵
2. *Classificazione matematica a soggetto MSC (Mathematics Subject Classification) e classificazione decimale Dewey CDD*, con due nuovi strumenti:
 - Connessione tra lo schema di classificazione decimale Dewey CDD21 e lo schema di classificazione matematica a soggetto MSC2000
 - Lista KWIC (KeyWords In Context) combinata per le descrizioni degli schemi CDD22 510 e MSC2000 + ZDM⁶ E-N
3. *Liste KWIC (KeyWords In Context) per le descrizioni di classificazioni scientifiche a soggetto*, con altri cinque nuovi strumenti:
 - Lista KWIC (KeyWords In Context) per le descrizioni dello schema MSC2000
 - Lista KWIC (KeyWords In Context) per le descrizioni dello schema ACM Computing Classification System (1998)
 - Lista KWIC (KeyWords In Context) per le descrizioni dello schema PACS 2001
 - Lista KWIC (KeyWords In Context) combinata per le descrizioni degli schemi MSC2000 e PACS 2001
 - Lista KWIC (KeyWords In Context) combinata per le descrizioni degli schemi MSC2000 e ACM Computing Classification System (1998)

Un passo successivo, ma già in fase di sviluppo all'interno del nostro gruppo di lavoro, è la messa a punto di una metodologia per l'interconnessione di classificazioni decisamente orientata a una rappresentazione concettuale. Tale metodologia verrà illustrata in questo articolo nell'ambito della presentazione di un'analisi relazionale della Mathematics Subject Classification, vista globalmente nel corso della sua evoluzione dal 1959, come è resa disponibile per le ricerche in linea nel database MathSci dell'American Mathematical Society.

La connessione di classificazioni nel mondo digitale

In un mondo sempre più interconnesso l'informazione digitale, bene immateriale astratto, diviene sempre più pervasiva. Diviene perciò una necessità, per le attività in rete basate sulla conoscenza, disporre di connessioni affidabili fra strumenti per la rappresentazione della conoscenza e strumenti lessicali. Strumenti per la ricerca di informazioni e documenti da una parte e strumenti lessicali quali schemi di classificazioni, soggettari, thesauri, terminologie e ontologie dall'altra, devono trovare il giusto equilibrio al fine di offrire servizi integrati all'utenza specializzata.

Utenti di diversi ambienti e con diverse richieste e aspettative desiderano soddisfare le loro necessità informative da qualunque parte l'informazione si trovi, risparmiando in termini di costi e tempi fin dove possibile.

Alla grande maggioranza degli utenti non interessano gli aspetti connessi alle caratteristiche specifiche delle diverse fonti, quanto il contenuto di testo, dati, immagini, ecc. dei documenti. L'utente però è costretto a muoversi in un contesto ove contenitori differenti seguono logiche e regole diverse: da basi di dati specializzate a cataloghi generali in linea di biblioteche, da

⁵ Antonella De Robbio, Dario Maguolo, Alberto Marini. Un progetto per la classificazione matematica, in "Biblioteche oggi", Maggio 2000 (Vol. 18, n. 4), p. 64-73

⁶ International Reviews on Mathematical Education. Zentralblatt fur Didaktik der Mathematik (ZdM)
<<http://www.mathematik.uni-osnabrueck.de/projects/zdm>>

motori di ricerca nel Web a portali dedicati, da basi di dati di riferimento (metadati) a biblioteche digitali di documenti a testo completo, da aggregatori di periodici elettronici a server di preprint e archivi personali degli autori (comunemente nominati e-print server).

L'organizzazione, le funzionalità e le modalità di interazione messe in atto dalle biblioteche digitali in rete sono completamente diverse da quelle in cui generalmente ci si imbatte nelle tradizionali biblioteche di documenti cartacei.

Gli utenti non gradiscono di dover reimpostare ragionamenti e modalità operative ogni volta che si trovano di fronte a diversi sistemi di registrazione, indicizzazione e presentazione delle informazioni: questo lavoro dovrebbe essere svolto automaticamente dal sistema, ma non è una cosa banale.

Sulla linea dei sistemi e-print server, Open Archives Initiative (OAI)⁷ sta promuovendo lo sviluppo di meccanismi tecnici e strutture organizzative per dare supporto alla loro interoperabilità. E' prevedibile che a breve ci sarà un'evoluzione che farà di questi sistemi i mattoni della costruzione di un modello del tutto nuovo di comunicazione scientifica, radicalmente differente dal modello tradizionale dominato dalla pesante mediazione delle aziende editoriali scientifico-accademiche, soprattutto per quanto concerne i periodici.

Per quanto riguarda l'indicizzazione semantica, succede spesso che differenti classificazioni, thesauri o terminologie altrimenti strutturate, pur interessando una stessa area, mantengano forti disaccordi sul piano linguistico (non si tratta di semplici questioni di traduzione), strutturale e semantico.

Si evidenziano disaccordi particolarmente gravi quando si passa dal mondo specialistico delle classificazioni disciplinari alle classificazioni generali di largo uso nelle biblioteche pubbliche, scolastiche e anche universitarie, quali la Classificazione Decimale Dewey, la Classificazione Decimale Universale, o la Classificazione della Library of Congress.

È facile incorrere in interpretazioni scorrette quando le stesse parole vengono usate in contesti o per scopi diversi. Ciononostante, è possibile realizzare connessioni effettivamente funzionanti fra classificazioni o strumenti simili, purché gli oggetti a cui ciascuna classificazione si riferisce siano identificati senza ambiguità per mezzo di un appropriato linguaggio di rappresentazione.

Classificazioni: funzioni, struttura e dinamica

Le classificazioni possono essere viste come spazi astratti strutturati, o modelli per organizzare spazi materiali, dove oggetti rispettivamente materiali o immateriali possono essere collocati, a seconda di caratteristiche selezionate, in modo che si possano trovare oggetti d'interesse semplicemente scegliendo e muovendosi nei percorsi predisposti dalla struttura dello spazio o dall'organizzazione definita dal modello. Tipici oggetti materiali che sono collocati tramite una classificazione sono i libri negli scaffali di una biblioteca; per quanto riguarda gli oggetti immateriali, possiamo pensare ai contenuti intellettuali, ai campi o alle discipline della conoscenza e dell'attività umana, ai concetti e agli oggetti di un certo campo o disciplina, o generalmente ai soggetti di documenti intesi in astratto come unità di informazione.

⁷ <<http://www.openarchives.org>>

Anche laddove una classificazione non è implicata nell'organizzazione di uno spazio fisico, per esempio in registrazioni bibliografiche trattate dal computer, si può dare per questa classificazione una semantica adeguata nei termini di una certa nozione di spazio, eventualmente meno vincolato e più complesso degli spazi materiali a cui siamo abituati. Lo spazio della classificazione è comunque la forma di un contenitore, uno spazio operativo, relativamente grossolano, per l'impacchettamento di concetti e l'inserimento dei pacchetti in una catena di trasporto.

Altra cosa è lo spazio dei concetti trasportati in un processo di comunicazione che si attua tramite una classificazione: uno spazio costituito dagli oggetti e dalle relazioni che sono effettivamente intesi dagli utenti della classificazione, sia per dichiarazione esplicita sia in modo parzialmente inconsapevole, ma sono comunque presenti e attivi nei processi interni di elaborazione dell'informazione di ciascun utente in forme e modi indipendenti, o almeno non totalmente condizionate, dai vincoli tecnici dei processi di comunicazione.

Visitando lo spazio di una classificazione, siamo aiutati da descrizioni strutturate, descrizioni testuali semistrustrate o non strutturate, immagini o altre forme di segnalazione, secondo specifiche convenzioni, a riconoscere luoghi e oggetti che vi trovano collocazione, o a prendere un'altra strada per spostarci da un luogo a un altro.

Ad ogni luogo, e quindi ad ogni oggetto, vengono assegnati uno o più indirizzi, che sono numeri o sequenze di caratteri funzionali a identificare il luogo tramite una codificazione del/dei percorsi per raggiungerlo.

Le descrizioni fanno riferimento agli oggetti attraverso la mediazione dei luoghi in cui gli oggetti sono raccolti, o dei canali che li veicolano, in funzione di un'adeguatezza rispetto a specificazioni o vincoli esterni (leggibilità da umani, trattabilità per usi specifici).

Perciò una singola descrizione può far riferimento a una collezione di oggetti che sono intesi distintamente dall'utente, ma sono raccolti in conformità all'organizzazione della classificazione.

D'altra parte, un singolo oggetto o luogo può essere rappresentato in forme differenti, pure nell'osservanza delle convenzioni linguistiche o semiotiche della classificazione.

I thesauri e i soggettari, al contrario, si preoccupano di mantenere una stretta corrispondenza fra oggetti e descrizioni, al prezzo di arrabattarsi su forme preferite e non preferite, ma questo significa costringere la varietà del linguaggio naturale a passare attraverso le ruote dentate degli identificatori di macchina. L'aggiunta di note d'ambito a testo libero o semistruurato è un ulteriore segnale di questa confusione.

È ruolo degli indirizzi la guida della macchina viaggiante: per svolgere questa funzione non c'è bisogno di conoscere perché la persona che viaggia vuole raggiungere un certo luogo, e per trovare che cosa. Nelle classificazioni di soggetto, quindi, gli indirizzi (comunemente chiamati codici o numeri di classificazione) sono fondamentali, nella loro forma esatta, per la collocazione di documenti materiali in biblioteche materiali. In tali spazi le liste di indirizzi sono i mezzi principalmente utilizzati per l'indicizzazione per soggetto nelle basi di dati bibliografiche e nei cataloghi in linea di biblioteche, ove gli indirizzi codificano e rendono intelleggibile la struttura dello spazio, ma agiscono nient'altro che come elementi di collegamento, senza nessun reale contenuto semantico. I reali vettori del contenuto semantico sono le descrizioni. La classificazione organizza questi vettori in una struttura che esiste indipendentemente dalle effettive forme degli indirizzi, cioè dalle forme che sono fissate allo scopo di rappresentare la struttura dello spazio della classificazione in vista di riferimenti e connessioni esterne.

Tanto le descrizioni quanto gli indirizzi possono cambiare, nel tempo o attraverso differenti convenzioni linguistiche, semiotiche o di codificazione; tuttavia non è necessario che descrizioni e indirizzi cambino in dipendenza le une con gli altri, o con i cambiamenti le trasformazioni, le nascite e le morti fra gli oggetti, gli spazi e i modi in cui oggetti e spazi sono organizzati e percepiti.

Differenti classificazioni che coprono aree in parte sovrapponibili possono esercitare nel tempo un'influenza reciproca, specialmente sulla struttura e sulle descrizioni, allo scopo di raggiungere per gli stessi oggetti modi di vedere simili o compatibili. Ciò avviene anche se i punti di vista o le estensioni di campo possono essere differenti, e ciascuna classificazione può continuare a raggruppare gli oggetti in maniera in tutto o in parte differente dalle altre

Visualizzazioni di schemi di classificazione: i nostri risultati

Mentre un certo numero di approcci alle questioni della connessione di classificazioni o thesauri utilizza metodi statistici o tecniche di reti neurali, una tendenza diversa è orientata all'analisi, modellazione e supporto dell'attività di organizzazione concettuale umana.

I primi possono essere di grande utilità, anche per queste ultime; un'integrazione ben definita sembra la ricetta per il prossimo futuro.

Il nostro lavoro non utilizza metodi statistici o tecniche di reti neurali.

Riguardo a questi problemi la nostra prima attenzione è stata rivolta alla generazione di ipertesti altamente portabili e a modalità di presentazione atte a facilitare la leggibilità e la scoperta del significato da parte di umani. In una generalità di complesse strutture di documentazione come schemi di classificazione, terminologie, collezioni di metadati, ecc. stiamo utilizzando con particolare attenzione una modalità di presentazione che permette di muoversi avanti e indietro attraverso viste parallele su una stessa struttura o su strutture simili. Questo si dimostra molto utile nel nostro ambito per varie ragioni, come vedremo nel seguito del presente lavoro.

Tali ipertesti sono prodotti principalmente da un gruppo di programmi in linguaggio C standard, che trattano solo file ASCII sequenziali e sono dedicati all'analisi e alla trasformazione di testi specifici e alla generazione di gruppi di pagine HTML sintatticamente semplici, ma altamente connesse e arricchite di procedure Javascript (*H-volumi*).

Sono state realizzate in questo modo diverse presentazioni ipertestuali a frame dell'ultima versione di Mathematics Subject Classification, MSC2000.

Queste presentazioni sono raccolte nella *Pagina della classificazione matematica*⁸, che costituisce la prima sezione del sito *La pagina delle classificazioni scientifiche*.

Ecco una lista dei gruppi di pagine HTML, insieme con l'indicazione delle loro caratteristiche e dell'indirizzo di accesso:

- ❑ Gruppo 1: Presentazione semplice a frame
- ❑ Gruppo 2: Presentazione a doppia visione
- ❑ Gruppo 3: Presentazione semplice a frame, con l'indicazione dei cambiamenti rispetto a MSC 1991

⁸ <<http://www.math.unipd.it/~biblio/math/index.html>>

un altro accesso è presso il server dell'Istituto IAMI-CNR di Milano:

<<http://www.iami.mi.cnr.it/~alberto/a0msc.htm>>

- ❑ Gruppo 4: Presentazione semplice a frame, con collegamenti a pagine specifiche di siti Web rilevanti
- ❑ Gruppo 5: Presentazione semplice a frame, traduzione italiana
- ❑ Gruppo 6: Presentazione semplice a frame, testi inglese e italiano interlineati

Siamo andati avanti su questa linea abbozzando delle connessioni fra i codici delle classificazioni Dewey e i codici della classificazione internazionale per la matematica MSC.

Il risultato è visibile dalla seconda sezione del sito "*Classificazione matematica a soggetto Mathematics Subject Classification MSC e classificazione decimale Dewey CDD*"⁹, che contiene due nuovi strumenti che coinvolgono le due classificazioni, la MSC2000 e la Dewey 21. ed.

E' stato quindi prodotto un primo strumento, una pagina a doppia visione "*Connessione tra lo schema di classificazione decimale Dewey CDD21 e lo schema di classificazione matematica a soggetto MSC2000*"¹⁰,

In vista della revisione della divisione 510 della CDD, *Matematica*, stiamo aggiornando questa bozza secondo la proposta presentata da Giles Martin, Assistant Editor della Classificazione Decimale Dewey¹¹.

Nel frattempo, abbiamo messo insieme le descrizioni di:

- la proposta di revisione della divisione 510 della CDD
- MSC2000
- le sezioni E - N della classificazione ZDM, codificate come 97E – 97N alla maniera della MSC

per produrre il secondo strumento di questa sezione, la *Lista KWIC (KeyWords In Context) combinata per le descrizioni degli schemi CDD22 510 e MSC2000 + ZDM E-N*¹².

Nella terza sezione del sito raccogliamo cinque H-volumi che presentano liste KWIC (Key-Word-In-Context) estratte dalle descrizioni di MSC2000, PACS 2001 e ACM-CCS 1998 da sole o in coppia (MSC PACS e MSC-CCS).

Le descrizioni sono permutate circolarmente, formando una lista KWIC, con perno su parole significative (filtrate da una lista di esclusione); la lista risultante è tanto pesante che, per mantenere una certa facilità di scorrimento, viene spezzata in file HTML circa 30 KB, che vengono visualizzati sulla destra, mentre sulla sinistra appare una lista alfabetica dei segmenti iniziali della prima stringa di ciascun file, che funziona da indice.

Questa presentazione, ridondante ma opportunamente impaginata, permette l'esplorazione rapida delle somiglianze lessicali fra le descrizioni, indirizzando all'individuazione delle affinità di contenuto. Proseguiremo nello sviluppo di questo tipo di supporto lessicale preliminare, investigando le connessioni fra diversi gruppi di schemi di classificazione.

Indagheremo inoltre su come si possano ottenere miglioramenti aggiungendo alle locuzioni la possibilità di discriminare omonimi, nonché sinonimi e termini secondari.

Bus nello spazio della classificazione

⁹ <<http://www.math.unipd.it/~biblio/math/itaddc.htm>>

¹⁰ <<http://www.math.unipd.it/~biblio/msc-cdd/index.html>>

¹¹ <http://www.oclc.org/dewey/updates/discussion/doc/request_for_comment.htm>

¹² <<http://www.math.unipd.it/~biblio/kwic/msc-cdd/index.html>>

Storicamente, le classificazioni di soggetto per documenti devono la loro struttura monogerarchica, ossia ad albero, a un'abitudine consolidata nel collocare a scaffale e ricercare documenti materiali nelle biblioteche, dove uno spazio fissato è definitivamente, o quasi permanentemente, diviso in sezioni e sottosezioni, e cambiare collocazione costa, tenendo presente che un oggetto materiale non può trovarsi simultaneamente in diversi luoghi. Tale abitudine consiste nel procedere a scelte per passare da argomenti generali a argomenti specifici, e quindi da grandi divisioni dello spazio a divisioni sempre più piccole, comunque senza sovrapposizioni. Anche nel mondo digitale questa abitudine è mantenuta e i sistemi per la gestione dei file presentano organizzazioni ad albero.

Tuttavia, gli spazi e gli oggetti della conoscenza e dell'attività umana mettono in evidenza strutture che sono molto più complesse e dinamiche di un semplice albero fisso. All'incrocio di intelligenza artificiale, linguistica computazionale e teoria delle basi di dati, queste strutture possono essere rappresentate con buona efficacia nel quadro di riferimento dell'Ontologia formale [G98], tramite formalismi come i grafi concettuali (CG)¹³, le Description Logic (DL)¹⁴, e lo Unified Modeling Language (UML)¹⁵, che viene dal campo dell'ingegneria del software e viene proposto come un approccio per la modellazione di ontologie e la codificazione del contenuto conoscitivo di pagine Web¹⁶.

Con questo orientamento si sta giungendo progressivamente alla definizione di formati di metadati per la rappresentazione di documenti.

La bozza dell'*Academic Metadata Format (AMF)*¹⁷, in via di definizione nell'ambito della Open Archives Initiative è un chiaro esempio di questa tendenza.

Come ci imbattiamo nelle classificazioni di soggetto, tali rappresentazioni strutturate (concepibili anche in tempi in cui mancavano linguaggi formali per esprimerle) devono essere drasticamente ridotte per avere qualche compatibilità con le forme ad albero in cui le classificazioni di soggetto vincolano la loro operabilità. Sebbene questa riduzione comporta inevitabilmente serie perdite di informazione, le classificazioni di soggetto sono state fornite di artifici più o meno efficaci per mettere rimedio a questa difficoltà.

Dal lavoro pionieristico di Ranganathan dal 1933 con la Classificazione Colon, attraverso le elaborazioni del Classification Research Group britannico negli anni '50 e '60, l'aggiunta delle Tavole Ausiliarie alla Classificazione Decimale Dewey dalla 18. ed., pubblicata nel 1971¹⁸, lo sviluppo del Preserved Context Indexing System (PRECIS) negli anni '70, e la pubblicazione nel 1986 dello standard ISO 2788 (BS 5723) *Guidelines for the establishment and development of monolingual thesauri*, si è progressivamente consolidato, con il nome di *analisi a facette*, un approccio compositivo all'analisi di soggetto¹⁹. Nell'analisi a faccette, i concetti complessi vengono decomposti in combinazioni specificate di elementi atomici, che appartengono a classi omogenee, reciprocamente esclusive, le *faccette*²⁰.

Ritornando alle classificazioni di soggetto, un'organizzazione dello spazio della classificazione (chiamata *precoordinazione*), la quale permette che oggetti complessi siano

¹³ S84; per un'applicazione di una variante of CG, ved. GMV99

¹⁴ JoLC99; CDLNR98; ved. *The DL Website* all'indirizzo:

<www.ida.liu.se/labs/iislab/people/patla/DL/index.html>

¹⁵ BJR98

¹⁶ C01

¹⁷ Thomas Krichel and Simeon Warner "Vocabulary for the Academic Metadata Format". Draft: 2001-03-26

<openlib.org/amf/doc/ebisu.html>

¹⁸ CCMS96

¹⁹ F96

²⁰ AG87

recuperati tramite indirizzi composti opportunamente, e un apparato più o meno organizzato di richiami e rinvii fra i luoghi, risultano mezzi utili specialmente se gli oggetti non possono essere collocati in più luoghi simultaneamente.

Se una classificazione di soggetto è usata in contesti che permettono l'uso simultaneo di differenti codici di classificazione per lo stesso oggetto, vengono forniti meccanismi e indicazioni per la *postcoordinazione*, allo scopo di recuperare parzialmente un significato complesso mettendo insieme in lista indirizzi in modi opportuni, sia nelle basi di dati che offrono informazioni, sia nelle interrogazioni che richiedono informazioni.

Senza illudersi di dare un taglio alle sfumature del linguaggio naturale, come i thesauri mostrano di fare, la base per identificare gli oggetti a cui una o più classificazioni fanno riferimento, nel contesto dell'organizzazione della conoscenza delle persone che usano quelle classificazioni, resta ancora un'analisi dei testi descrittivi, insieme con la ricognizione delle dinamiche strutturali nel tempo delle stesse classificazioni. Segnali importanti per queste operazioni di identificazione sono dati da identità o somiglianze nel testo delle descrizioni in una classificazione (scoperte dalla stessa versione – sincronicamente – o da diverse versioni nel tempo – diacronicamente –), o attraverso diverse classificazioni (nella stessa lingua), sebbene una parola o locuzione può avere significati diversi, anche all'interno della stessa classificazione.

Perciò il primo passo nel processo di tirar fuori gli oggetti dalla loro busta, cioè lo spazio della classificazione, è di riconoscere la busta come una struttura che si sviluppa nel tempo attraverso una sequenza di versioni successive della classificazione, muovendosi fra gli indirizzi che segnano i percorsi e i luoghi in tale spazio. Il prodotto di questa fase consiste di sequenze di descrizioni per i *bus* che corrono nello spazio-tempo della classificazione. Ogni bus durante la sua corsa attraversa uno o più luoghi; gli indirizzi di questi luoghi, insieme con l'indicazione del periodo di passaggio, costituiscono la tabella di corsa di quel bus.

Per produrre dei risultati su questa linea guida, abbiamo iniziato un'analisi dell'intero corpus della Mathematics Subject Classification, in tutto il corso della sua evoluzione dal 1959, così come è disponibile per le ricerche in linea nel database *MathSci*. Ne è venuto fuori un database relazionale, per il quale stiamo specificando una presentazione Web da realizzare con un adattamento dei programmi C già sviluppati.

L'identificazione e la descrizione degli oggetti fuori dello spazio dei bus

Anche se ogni sezione sincronica dello spazio dei bus è ad albero, l'intera struttura può non essere ad albero, dato che nodi o sottoalberi possono migrare da un ramo a un altro; oltre alla struttura gerarchica principale, richiami e rinvii, e i meccanismi esplicitamente dichiarati di preordinazione e postordinazione, anch'essi considerati dinamicamente, danno contributi sostanziali alla definizione dello spazio-tempo della classificazione.

Il passo ulteriore del processo di identificazione degli oggetti è l'estrazione di elementi concettuali dalle descrizioni. Si dovrebbe quindi attuare un'approfondita analisi relazionale per mezzo di un linguaggio di rappresentazione adeguato.

Un aiuto efficace in questa fase viene dato da strumenti per l'analisi testuale, quali le visualizzazioni di liste KWIC che abbiamo illustrato più sopra.

Le descrizioni degli oggetti nella macchina dei metadati

Mathematics Subject Classification è uno dei sistemi di classificazione previsti dal formato di metadati Dublin Core (DC), ed è utilizzata nei metadati DC per il motore di ricerca sviluppato nel progetto dell'Unione Europea *European Libraries and Electronic Resources in Mathematical Science* (EULER)²¹.

L'obiettivo principale di EULER è stato la realizzazione di un "one-stop shop" per la ricerca di risorse informative per la matematica quali libri, preprint, pagine Web, abstract, collezioni di articoli e recensioni, periodici, rapporti tecnici e tesi. Il risultato è una meta-interfaccia Web per interrogazioni parallele dirette a una collezione eterogenea di database.

Una strategia simile potrebbe essere messa in campo per la connessione delle classificazioni: le descrizioni degli oggetti identificati a partire da diverse classificazioni, opportunamente codificate, potrebbero entrare a far parte dei metadati gestiti dal motore di ricerca.

Il modo più ovvio ed efficace di interconnettere classificazioni di soggetto, tesauri o liste di intestazioni di soggetto è comunque fornito dagli stessi record bibliografici, laddove più di un sistema è utilizzato per l'indicizzazione semantica in uno stesso record.

In effetti, all'interno di cataloghi o di strumenti bibliografici diversi, lo stesso documento viene ad essere prevalentemente rappresentato con indici tratti da differenti sistemi.

Mentre gli OPAC di biblioteche generali si basano fondamentalmente sullo schema Dewey per la classificazione e sulle liste nazionali di intestazioni di soggetto per la soggettazione, i database bibliografici specializzati sono dotati di sistemi di indicizzazione specifici, quali schemi di classificazione disciplinare o tesauri specializzati.

Sarebbe sufficiente, come approccio di base, offrire uno strumento di raccordo, semplice ma efficace, mettendo questi record assieme al fine di un matching per creare il collegamento.

In questo modo, scorrendo all'interno della struttura di un sistema di indicizzazione semantica si può giungere sia al metadato del documento, sia ad un accesso diretto al testo completo del documento stesso, oppure si può passare da un sistema di indicizzazione ad un altro per esplorazioni differenti e ulteriori navigazioni tra sistemi diversi che connettono ad altri metadati o documenti.

Al fine di monitorare la navigazione tra schemi classificatori differenti sarebbero necessari appositi metadati per l'identificazione delle versioni dei sistemi di indicizzazione semantica.

Un formato di metadati per tali oggetti ad oggi non è ancora stato approntato.

In questa prospettiva occorre iniziare un lavoro di codifica degli schemi delle classificazioni scientifiche esistenti, procedendo ad una loro dettagliata descrizione in relazione alle versioni esistenti, all'interno di formati di metadati che prevedano il loro accoglimento entro strutture di tipo "nesting".

Un altro approccio pragmatico agli strumenti da noi sviluppati potrebbe essere l'utilizzo degli schemi di classificazione interconnessi e delle liste KWIC visti come portali gateway di snodo ai documenti contenuti nei preprint server. Attraverso le liste KWIC attualmente disponibili, le quali contengono i codici classificatori di ciascuno schema, mondi diversi potrebbero colloquiare in differita. Sarebbe interessante poter offrire un servizio in un'ottica di service provider OAI compatibile ove, data una locuzione verbale presente nella lista KWIC e recuperato il codice che rimanda allo schema classificatorio specialistico di pertinenza, ci si

²¹ <<http://www.emis.de/projects/EULER/>>

connetta ai metadati esposti dai data provider, di modo da poter raggiungere i documenti a testo pieno presenti sui singoli server disciplinari.

Per il prossimo futuro, le parole chiave che indicizzeranno i futuri sviluppi di un lavoro coordinato sulle classificazioni scientifiche saranno OPAC, e-print server OAI e metadati.

RIFERIMENTI BIBLIOGRAFICI

- AG87 J. Aitchison, A. Gilchrist, "Thesaurus construction: a practical manual", ASLIB, 1987
- BJR98 G. Booch, L. Jacobson, J. Rumbaugh, "The Unified Modeling Language User Guide", Addison-Wesley, 1998
- C01 S. Cranefield, *Networked Knowledge Representation and Exchange using UML and RDF*, "Journal of Digital Information", 1(8), 2001, <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Cranefield/>
- CCMS96 L.M. Chan, J.P. Comaromi, J.S. Mitchell, M.P. Satija, "Dewey Decimal Classification: a practical guide. 2nd ed., revised for DDC 21", OCLC Online Computer Library Center, 1996
- CDLNR98 D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati, *Description logic framework for information integration*, "Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98)", Morgan Kaufman, 1998. p. 2-13
- D01 M. Doerr, *Semantic Problems of Thesaurus Mapping*, "Journal of Digital Information", 1(8), 2001, <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/>
- F96 A.C. Foskett, "The Subject Approach to Information", 5th ed., Library Association Publishing, 1996
- G98b N. Guarino, *Formal Ontology and Information Systems*, "Formal Ontology in Information Systems: proceedings of FOIS'98", N. Guarino (ed), IOS Press, 1998. p 3-15
- GMV99 N. Guarino, C. Masolo, G. Vetere, *Ontoseek: Content-based Access to the Web*, "IEEE Intelligent Systems", 14(3), 1999. p. 70-80
- H01 J. Hunter, *MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains*, "Journal of Digital Information", 1(8), 2001, <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/>
- JoLC99 "Journal of Logic and Computation" – Vol. 9, No. 3: "Special Issue on Description Logics"
- S84 J. Sowa, "Conceptual Structures: Information Processing in Minds and Machines", Addison-Wesley, 1984